JAIA – Journal Of Artificial Intelligence And Applications
Journal Homepage : http://jurnal.sar.ac.id/index.php/JAIA/

Sentiment Analysis to analyze Vaccine Enthusiasm in Indonesia on Twitter Social Media

M. Khairul Anam[1], Rahmaddeni[2], Muhammad Bambang Firdaus[3], Hadi Asnal[4], Hamdani[5]
[1,2,4,5]STMIK Amik Riau, JL. Purwodadi Indah Km 10 Panam, Pekanbaru, Indonesia
[3]Mulawarman University, Jalan Sambaliung No.9 Sempaja Selatan Samarinda Utara, Samarinda, Indonesia

Article Info

Received : Apr 10, 2021
Revised : Apr 28, 2021
Accepted : Apr 30, 2021

Keywords :

Vaksin
Tweet
Sentimen Analysis
Naïve bayes

Abstract

Vaccines are one way to prevent the coronavirus from entering the human body, although it is not 100% accurate. However, the implementation of vaccination in Indonesia is still controversial. People give their opinions directly or through social media such as Twitter. Twitter is one of the most frequently used social media as a dataset in data mining research. To take tweets as data from Twitter is mostly done in various ways, for example, using an API that is connected to other tools such as Python or Rapidminer. In addition, you can also use the Drone Emprit Academy portal as a tool. The data obtained is then preprocessed using case folding, cleaning, tokenizing, filtering, and stemming. After that, the model was evaluated using the Naïve Bayes method. Naïve Bayes is a classification method that can predict the probability of a class, so that it can produce decisions based on learning data. Currently, Naïve Bayes is one of the best methods to find accuracy in sentiment analysis that is often to used. The results of this study obtained an accuracy of 79%.

1. Introduction

Twitter is one social media with the most significant users and has a very diverse age range [1]. Tweets on Twitter are widely used for research such as sentiment analysis [2], Social Network Analysis (SNA) [3][4], classification [5], clustering [6], and so on. Twitter also has a trending topic feature that people use as news in various media such as television, news portals, and so on [7].

News related to vaccination is being talked about a lot. Many people accept and reject the COVID-19 vaccine [8]. Acceptance and rejection of vaccines on social media is also a form of online participation from the public towards the government [9]. Several vaccines prevent covid 19 in Indonesia, such as Astra Zeneca, Sinopharm, Moderna, Pfizer-BioNTech, and Sinovac Biotech Ltd [10][11]. Due to the many pros and cons that occur, this study will conduct sentiment analysis to see positive, negative, and neutral tweets.

Sentiment analysis is the process of automatically extracting, understanding, and processing data in the form of unstructured text to obtain sentiment information contained in a sentence of opinion or opinion [12]. Previous researchers have discussed several studies related to sentiment analysis. Sentiment analysis is carried out to see the accuracy generated by several methods such as Support vector machine (SVM) [13],

M. Khairul Anam
Email : khairulanam@sar.ac.id

Naïve Bayes, K-Nearest Neighbor [14], C.45 Algorithm, Random Forest, Decision Tree [15], and so on.

Research conducted by Franciska and arief [16] analyzing the customer sentiment of the online store jd.id using the nave Bayes classifier to get an accuracy of 96.44%. Then research billy, helen, dan enda [17] conducted a product analysis using the Naive Bayes method to get an accuracy of 90%. Another study [18] which also used the naive Bayes classifier, got an accuracy of 70%.

This study will also use the naive Bayes classifier method to obtain high accuracy in conducting sentiment analysis on enthusiasm for the Covid-19 vaccination in Indonesia.

2. Research Methods

The research uses a methodology flow to make it easier to carry out the research process.

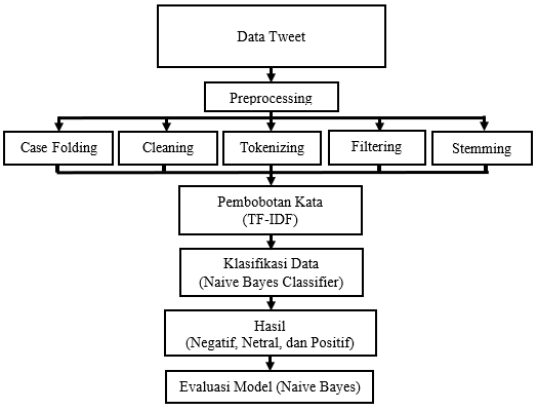

Figure 1. Research methodology flow

The preprocessing of this research will use several stages, namely case folding is used to convert all letters in the document into lowercase letters; only letters 'a' to 'Z are accepted. Characters other than these letters are omitted and are considered delimiters [19]. Then cleaning is used for delimiter or deletion of characters or punctuation marks and URLs and emoticons [20]. Tokenizing is a process to make a sentence more meaningful by breaking the sentence into words [21]. Only take words that have an important meaning in the training data [22]. Furthermore, finally stemming, stemming is a process that provides a mapping between various words with different morphology into one basic form (stem) [23]. After preprocessing, the next step is word weighting.

Weighting is converting words into numbers (word vector) [24]. The weighting is done by Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF is a method that aims to give weight to the relationship of a word (term) to a document or comment [25]. After weighing the words, the next step is to carry out a sentiment analysis process, and the last step is to evaluate the model.

3. Results and Discussion

The weighting process is carried out after the preprocessing process. In this research, the process of making word vectors and word weighting uses the help of the Python3 library, namely TfidfVectorizer. The vector representation results obtained 700 numbers which have 2153 words. The results of the TF-IDF word weighting can be seen in Figure 2.

```
700 Number of reviews has 2153 words
  (0, 795)      0.3697612450197021
  (0, 1729)     0.3697612450197021
  (0, 1875)     0.36095470381707717
  (0, 1340)     0.35318678395874953
  (0, 1292)     0.11544010027963734
  (0, 1494)     0.09161880126751445
  (0, 2076)     0.13227904828145906
  (0, 2039)     0.5770000466860385
  (0, 396)      0.3152408901204811
  (1, 63)       0.34937995202670064
  (1, 183)      0.25238352128791863
  (1, 80)       0.3035117037257312
  (1, 960)      0.311856610503672
  (1, 387)      0.27000017791326614
  (1, 717)      0.3035117037257312
  (1, 600)      0.3217265403460804
  (1, 1198)     0.3035117037257312
  (1, 517)      0.2229550411156791
  (1, 935)      0.18631195864170447
  (1, 2078)     0.35618795959403693
  (1, 2110)     0.21052708869751274
  (1, 1494)     0.07520360459769337
  (1, 2076)     0.05428940952016844
  (2, 1382)     0.4483043690068737

show more (open the raw output data in a text editor) ...

  (699, 1332)   0.6596854824792299
  (699, 1857)   0.45539520557669205
  (699, 948)    0.5256600323661662
  (699, 1292)   0.1872634712812819
  (699, 2076)   0.21457910811724853
```

Figure 2. Word weighting

After going through data preprocessing and weighting, a model is then made that will

be used to classify the test data. This process is carried out using a Python 3 programming language library called sci-kit-learn for the classification process.

The sentiment analysis results are in the form of negative, neutral, and positive opinion categories. As for the details, see Table 1.

Table 1. Opinion Category Results

Category	Result
Negative	118
Neutral	15
Positive	217

The results showed that the "Positive" category was neutral and negative. The data is visualized in a bar chart and word cloud as follows.

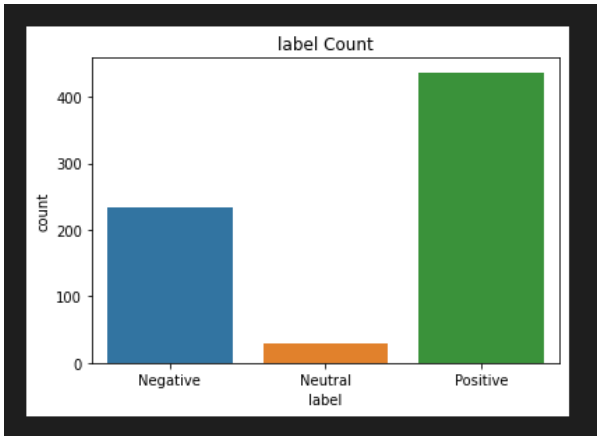

Figure 3. Opinion category diagram results

Based on Figure 4 is the overall word cloud visualization of the imported data sources, the word 'vaccine' is the word with the most frequency, followed by the words 'Pfizer, 'moderna', 'covid,' 'booster,' 'vaccination' and 'Sinovac's.

Gambar 4. Whole wordcloud

After the sentiment analysis process, the next step is to evaluate the model. Figure 5 is an evaluation of the model.

	precision	recall	f1-score	support
Negative	0.85	0.52	0.65	21
Neutral	0.00	0.00	0.00	3
Positive	0.77	0.96	0.85	46
accuracy			0.79	70
macro avg	0.54	0.49	0.50	70
weighted avg	0.76	0.79	0.76	70

Figure 5. NBC model evaluation

From Figure 5, it can be seen that the accuracy of this research is 79%.

4. Conclusion

Results of the research that has been done show that there are 700 tweets on Twitter and have 2153 words. The results of the model's evaluation obtained an accuracy of 79%. The results prove that the nave Bayes method is still one of the best. Although this research does not use feature selection, nave Bayes can still produce the best accuracy.

5. Reference

[1] G. Appel, L. Grewal, R. Hadi, and A. T. Stephen, "The future of social media in marketing," *J. Acad. Mark. Sci.*, vol. 48, pp. 79–95, 20219, doi: 10.1007/s11747-019-00695-1.

[2] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," in *Procedia Computer Science*, 2019, vol. 161, pp. 765–772, doi: 10.1016/j.procs.2019.11.181.

[3] I. Febrianti, M. K. Anam, Rahmiati, and Tashid, "Tren Milenial Memilih Jurusan Di Perguruan Tinggi Menggunakan Metode Social Network Analysis," *Techo.COM*, vol. 19, no. 3, pp. 216–226, 2020, doi: https://doi.org/10.33633/tc.v19i3.3483.

[4] M. K. Anam, T. P. Lestari, Latifah,

M. B. Firdaus, and S. Fadli, "Analisis Kesiapan Masyarakat Pada Penerapan Smart City di Sosial Media Menggunakan SNA," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 69–81, 2021, doi: https://doi.org/10.29207/resti.v5i1.2742.

[5] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach for bot detection on Twitter," *Comput. Secur.*, vol. 91, pp. 1–14, 2020, doi: 10.1016/j.cose.2020.101715.

[6] F. E. Ayo, O. Folorunso, F. T. Ibharalu, I. A. Osinuga, and A. Abayomi-Alli, "A probabilistic clustering model for hate speech classification in twitter," *Expert Syst. Appl.*, vol. 173, no. February, p. 114762, 2021, doi: 10.1016/j.eswa.2021.114762.

[7] H. U. Khan, S. Nasir, K. Nasim, D. Shabbir, and A. Mahmood, "Twitter trends: A ranking algorithm analysis on real time data," *Expert Syst. Appl.*, vol. 164, no. September 2020, p. 113990, 2021, doi: 10.1016/j.eswa.2020.113990.

[8] T. El-Elimat, M. M. AbuAlSamen, B. A. Almomani, N. A. Al-Sawalha, and F. Q. Alali, "Acceptance and attitudes toward COVID-19 vaccines: A cross-sectional study from Jordan," *PLoS ONE*, vol. 16, no. 4 April. 2021, doi: 10.1371/journal.pone.0250555.

[9] M. K. Anam, "Analisis Respons Netizen Terhadap Berita Politik Di Media Online," *J. Ilm. Ilmu Komput.*, vol. 3, no. 1, pp. 14–21, 2017, doi: 10.35329/jiik.v3i1.62.

[10] R. N. Rahayu and Sensusiyati, "Vaksin covid 19 di indonesia : analisis berita hoax," *Intelektiva J. Ekon. Sos. Hum. Vaksin*, vol. 2, no. 07, pp. 39–49, 2021.

[11] P. J. Turner *et al.*, "COVID-19 vaccine-associated anaphylaxis: A statement of the World Allergy Organization Anaphylaxis Committee," *World Allergy Organ. J.*, vol. 14, no. 2, p. 100517, 2021, doi: 10.1016/j.waojou.2021.100517.

[12] T. Le, "An attention-based deep learning method for sentiment analysis," in *Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020*, 2020, pp. 282–286, doi: 10.1109/CSCI51800.2020.00054.

[13] M. Demircan, A. Seller, F. Abut, and M. F. Akay, "Developing Turkish sentiment analysis models using machine learning and e-commerce data," *International Journal of Cognitive Computing in Engineering*, vol. 2. pp. 202–207, 2021, doi: 10.1016/j.ijcce.2021.11.003.

[14] R. N. Devita, H. W. Herwanto, and A. P. Wibawa, "Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, p. 427, 2018, doi: 10.25126/jtiik.201854773.

[15] W. Baswardono, D. Kurniadi, A. Mulyani, and D. M. Arifin, "Comparative analysis of decision tree algorithms: Random forest and C4.5 for airlines customer satisfaction classification," in *Journal of Physics: Conference Series*, 2019, vol. 1402, no. 6, doi: 10.1088/1742-6596/1402/6/066055.

[16] F. V. Sari and A. Wibowo, "Analisis Sentimen Pelanggan Toko Online Jd. Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 2, no. 2, pp. 681–686, 2019, doi: https://doi.org/10.24176/simet.v10i2.3

487.

[17] B. Gunawan, H. S. Pratiwi, and E. E. Pratama, "Sistem Analisis Sentimen pada Ulasan Produk Menggunakan Metode Naive Bayes," *J. Edukasi dan Penelit. Inform.*, vol. 4, no. 2, p. 113, 2018, doi: 10.26418/jp.v4i2.27526.

[18] V. A. Permadi, "Analisis Sentimen Menggunakan Algoritma Naive Bayes Terhadap Review Restoran di Singapura," *J. Buana Inform.*, vol. 11, no. 2, p. 140, 2020, doi: 10.24002/jbi.v11i2.3769.

[19] G. P. A. Brahmantha and I. W. Santiyasa, "Sentiment Analysis of the Enforcement of PSBB Part II in Jakarta," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 9, no. 2, p. 259, 2020, doi: 10.24843/jlk.2020.v09.i02.p13.

[20] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[21] H. M. Zin, N. Mustapha, M. A. A. Murad, and N. M. Sharef, "The effects of pre-processing strategies in sentiment analysis of online movie reviews," in *AIP Conference Proceedings*, 2017, vol. 1891, doi: 10.1063/1.5005422.

[22] W. Gata and Purnomo, "Akurasi Text Mining Menggunakan Algoritma K-Nearest Neighbour pada Data Content Berita SMS," *J. Format*, vol. 6, no. 1, pp. 1–13, 2017.

[23] N. Z. Dina and N. Juniarta, "Aspect based Sentiment Analysis of Employee's Review Experience," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 1, p. 79, 2020, doi: 10.20473/jisebi.6.1.79-88.

[24] M. N. Saadah, R. W. Atmagi, D. S. Rahayu, and A. Z. Arifin, "Information Retrieval Of Text Document With Weighting Tf-Idf And Lcs," *J. Comput. Sci. Inf.*, vol. 6, no. 1, pp. 34–37, 2013, doi: https://doi.org/10.21609/jiki.v6i1.216.

[25] M. N. Saadah, R. W. Atmagi, D. S. Rahayu, and A. Z. Arifin, "Sistem Temu Kembali Dokumen Teks Dengan Pembobotan Tf-Idf Dan Lcs," *JUTI J. Ilm. Teknol. Inf.*, vol. 11, no. 1, p. 19, 2013, doi: 10.12962/j24068535.v11i1.a16.